DOCUMENT RESUME

ED 095 000 SE 017 719

AUTHOR Herron, J. Dudley; And Others
TITLE The Proper Experimental Unit: Comparative Analyses of Empirical Data.
PUB DATE Apr 74
NOTE 16p.; Paper presented at the annual meeting of the National Association for Research in Science Teaching (47th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Chemistry; College Science; College Students; Educational Objectives; *Educational Research; *Research Design; Research Problems; Science Education
IDENTIFIERS Research Reports

ABSTRACT

Reported is a discussion concerning the effect of different analyses of the same empirical data. Students (N=over 200) enrolled in an introductory college chemistry course for science majors were randomly assigned to two experimental treatments. Treatment O was designed to help students understand the objectives of the course and to emphasize the importance of the objectives. Students in treatment R also received a list of objectives for the course but emphasis was on providing feedback concerning their progress toward meeting these objectives via a weekly 10 point quiz. Half the students in each class in treatment R were told they had to score at least 8 points on the quiz or they would receive a zero grade for that quiz. They could, however, re-take the quiz as often as they wished in order to achieve a score above 8 points. The other students in the class could also re-take the quizzes but they were under no coercion concerning their scores. For treatment O, the appropriate experimental unit is the class section but the appropriate experimental unit for treatment R is not clear. The opportunity to re-take a quiz appeared to indicate that the appropriate experimental unit, for R, was the individual. (PEB)

ED 095000

THE PROPER EXPERIMENTAL UNIT: COMPARATIVE ANALYSES OF EMPIRICAL DATA

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

By
J. Dudley Herron*
Tom Luce
Van E. Neie

Purdue University
W. Lafayette, Indiana

INTRODUCTION

If there is such a thing as a "typical" study in education it would be described as a study in which two or more existing classrooms are selected from some convenient population of possible classes (usually classrooms within a suitable driving distance of a particular university), demographic data are obtained which suggest that these classes are reasonably comparable, some kind of experimental treatment is administered to half of these classes (usually, but not always, selected more or less at random), and the performance of the "experimental" and "control" groups are compared on some criterion measure.

In such a design, the treatment is assigned at random to classes (or perhaps even to groups of classes taught by the same teacher) rather than to individuals within classes. It has been pointed out that the appropriate experimental unit in such studies is the class rather than the individual and that the data should be analyzed using class means as the raw data rather than individual scores on the criterion measure. (Raths, 1967). Still, most data analyses in science education studies are based on the use of the individual as the experimental unit.[1] There are several possible explanations for this. First, many

*Inquiries regarding this paper should be directed to Dr. J. D. Herron, Dept. of Chemistry, Purdue University, W. Lafayette, Indiana 47907

[1] In Vol.10, Nos.1-3 of JRST there are 10 studies for which there may be a question concerning the appropriate experimental unit. Of these, 8 used the individual while 2 used the class.

SE 017 719

researchers are simply ignorant of the theoretical arguments for using the class as the experimental unit. Second, these experimenters want to obtain significant results and feel that the reduction in degrees of freedom which result from using the class rather than the individual as the experimental unit will reduce the chance of getting these differences. Third, some researchers, perhaps bolstered by the arguments of Fletcher (Fletcher, 1968), are not convinced that using group means as the experimental unit is a better procedure than using the individual score.

Whatever the reasons and regardless of the soundness of the arguments, the reader of research is faced with the fact that many of the studies which may interest him have been conducted under circumstances for which there is some question concerning the appropriateness of the experimental unit used in the analysis. What does one do? Should one ignore results and conclusions based on analyses in which the wrong experimental unit was used in the analysis? Should the results be accepted without question? These are the issues with which we are concerned.

We do not claim that we have the answers to the questions that we have raised. However, we do have data, based on one empirical study, which we think shed some light on these questions. We believe these data suggest that the problem may be less serious than some would suggest, that interpretations of data are likely to be similar (though certainly not identical) irrespective of the treatment, and that one may err more by dismissing a study out of hand because an incorrect choice of experimental unit was made rather than accepting the results as "probably correct."

THE NATURE OF THE STUDY

The concern of this paper is with the effect of different analyses of the

same empirical data. But it is necessary, first, to describe the nature of the study on which these analyses were performed.

The study grew out of a concern over the effect of providing behavioral objectives to college students. An earlier study had suggested that it made little difference whether students in a beginning college chemistry course were given lists of behavioral objectives, (Herron, 1971). Those students who were given the lists seemed to do little better than those who did not have them. This result, being contrary to popular educational bias, started a search for the reason.

One explanation entertained by the authors was that the students did not really understand what the lists were saying.[2] It was decided that a treatment which would "explain" the objectives to students should be included in the study under discussion here.

Another possible explanation for the earlier observation that lists of objectives had little effect on college chemistry students was that the students simply did not make use of what they had - that they needed a little "coercion" to get them to use the objectives. It was decided to provide some form of "coercion" in this study.

The study was conducted in an introductory college chemistry course for science majors. Over 200 students were enrolled in the course. All students met for a large lecture session twice a week and met once for recitation and once for a three-hour laboratory. Students had the same graduate assistant for lab and recitation. There were a total of twelve of these small sections in the course with enrollment ranging from a low of 14 to a high of 22, with a mean enrollment of 18. The two experimental treatments were assigned at random to

---

[2] This hypothesis grew out of some unpublished work by Herron and Hiscox which suggested that students had difficulty in matching objectives with test items over the objectives.

class sections, so that there were three sections represented in each of the four cells of the 2 x 2 factorial design.

The two treatments in the design consisted of the following:

1. Treatment O: This treatment was intended to help students understand the objectives of the course and to emphasize the importance of the objectives. Students in a class section receiving this treatment found that each of their recitation classes was organized around the list of objectives for the week. The class was conducted by going over the list of objectives, trying to determine which objectives the students were having difficulty with and providing help with these objectives. In most instances, the help would be in the form of a referral to one or more of the homework problems which were related to the objective.

   If a class section did not receive Treatment O, the recitation session was used for studying assigned homework. Objectives were not mentioned unless a student asked a specific question about them. It should be emphasized that <u>all</u> students had a list of objectives for the course. These were handed out in the laboratory each week. Treatment O simply represents a difference in the attention given to the lists of objectives during the recitation session.

2. Treatment R: A quiz was given to all students during the first half-hour of each laboratory session. This quiz was related to the objectives for the previous week and was scored on a 10 point basis. The primary purpose of the quiz was to provide students with feedback concerning their progress toward meeting the objectives of the course. In order to provide a "coercion" treatment for some individuals in the course, students in half the class sections were told that they must either score at least 8 points on the weekly quiz or receive a grade of zero for that quiz. If they scored below 8 points they had an opportunity to re-take the quiz as many times as they liked but

their score would remain zero until they scored above 8 points; thus, possible scores for students in this treatment were 10, 9, 8 and 0. Students in other class sections were given the same weekly quiz and the same opportunity to re-take the quiz as many times as they wished. However, there was no attempt to coerce them to do so. These students received whatever score they made on the quiz at the first administration or, if they took the quiz over, the score that they received the second time, be it higher or lower. Thus, a student under the "0" treatment who scored below 8 points - a 7, for example - had nothing to lose by re-taking the quiz and 8 points to gain. A student who was <u>not</u> in the treatment group and who scored 7 points had little to gain (a maximum of 3 points) by re-taking the quiz and he could possibly lose since his final score for the quiz would be that which he obtained on the last administration of the quiz.

In summary, the basic 2 x 2 factorial design consisted of random assignment of class sections to each of four cells represented in Table I.

<u>Table I</u>

<u>Treatment 0*</u>

Treatment R*	0	o
R	1c 3b 6b	1b 4a 6a
r	1a 2a 4b	2b 3a 6c

*Upper case letters represent cells containing class sections that received the indicated treatment; lower case letters represent those cells for which the indicated treatment was absent.

THE EXPERIMENTAL UNIT

Now that the design of the study has been briefly outlined, we turn to the question of the appropriate experimental unit. It seems clear that for Treatment 0, the appropriate experimental unit is the class section. The treatment is administered to the section as a whole and it is very likely that interactions among students within the class constitute an important part of the treatment. Thus, the appropriate analysis of the data would be to treat the means for each class section as a "score" for that section and conduct the analysis of variance accordingly.

The appropriate experimental unit for Treatment R is not as clear. Although Treatment R was assigned to class sections at random rather than to individuals, the treatment itself is essentially an individual treatment. Each individual student decided whether he wanted to repeat a quiz or receive a grade of zero. There is little reason to believe that the choices of others in his class section would have any important influence on his decision since everyone who repeated the quiz did so individually and outside of class time. Although the reader may disagree, we are inclined to say that the appropriate experimental unit for Treatment R is the individual.

Partially as a result of the "mixed" nature of our experiment, but primarily out of curiosity, it was decided that the data from this study would be analyzed in several ways. These various analyses are summarized in Table II. The simplest of these analyses (represented by Ia in Table II) consists of a conventional two-way ANOVA using the individual as the experimental unit. The second analysis in the table (Ib) is identical to the first with the exception that the section mean is used as the experimental unit. These two analyses, using various measures as the criterion variable, provide an opportunity to compare the conclusions that would result from the same data when either the individual or the class is considered as the experimental unit.

Table II*

Types of Analyses Performed

Type of Analysis	Criterion Measure	Experimental Unit	3rd Factor	Co-Variates
Ia	↑ TP or Q08 ↓	Individual Score	---	---
Ib		Section Mean	---	---
IIa		Individual Score	---	V-M
IIb		Section Mean	---	V-M (mean)
IIIa		Individual Score	M	---
IIIb		Split Section Mean	M (rm)	---
IVa		Individual	M	Qrep
IVb		Split Section Mean	M (rm)	Qrep

*For an explanation of the symbols used in this table, refer to the legend accompanying Table VIII, page 14.

Since there is always the question of the comparability of class groups, other analyses were performed in which SAT-M and SAT-V scores were used either as covariates or as a third factor in the factorial design. For example, analysis IIa in Table II is identical to analysis Ia with the exception that the verbal and math scores on the SAT exam have been used as covariates to statistically adjust the scores on the criterion measure for differences in ability that might have existed between treatment groups. Analysis IIb parallels Ib and is equivalent to analysis IIa with the exception that the class is treated as the experimental unit. The scores used for the covariate adjustment are the mean SAT-V and the mean SAT-M for the class section. The next pair of analyses shown in Table II (IIIa and IIIb) represent those for which the SAT scores are used as a third factor in a 2 x 2 x 3 factorial design. In analysis IIIa, the sample of students was stratified into a high, average, and low SAT group and the ANOVA was done to determine if there was a main effect due to "ability" as measured by

the SAT-M test. It should be noted that analysis IIIb is not exactly parallel to IIIa. Since the section means represent contributions from high, average, and low ability students, an analysis in which there is a stratification of SAT section means does not make any sense. The analysis which was done is one suggested by Page, (Raths, 1967), in which one stratifies each section into a high, average and low ability (SAT in our case) groups calculates both the mean SAT and the mean on the criterion measure, and then treats the data as though they are repeated measures of the section mean on the criterion measure when the section has a high SAT, when the same section has an average SAT, and when the same section has a low SAT score. This appears to be the nearest equivalent to analysis IIIa when one wishes to use the class as the experimental unit.

Still a fourth kind of analysis is represented by analyses IVa and IVb. This pair is identical to IIIa and IIIb with the exception that now both a covariate and a third factor are added in the analysis.

We have summarized the kinds of analyses which were performed. There were, in fact, several analyses of each type carried out. These differed in what was used as the criterion measure. A total of over forty analyses of variance and covariance were performed of which 24 are presented in this paper. In the analysis which we are presenting, only two criterion measures are used. One is the total number of points (see legend for Table VIII) that the student accumulated in the course and the second is the number of quizzes on which the student scored eight or above. The rationale for choosing these as criterion measures was quite simple. It was assumed that the total points accumulated in the course was likely to be the most sensitive measure of student achievement in the course and we were primarily interested in knowing how our treatments would affect achievement. We first selected "number of quizzes over eight" as a criterion measure to see if we did in fact have an "R" treatment. If our admonition to repeat quizzes on which a score of less than eight points was obtained was taken seriously, then

we should certainly see a "main effect" for Treatment R when "number of quizzes over eight" is taken as the criterion measure. Later, this criterion measure proved useful in comparing the interpretations that would be drawn from the results of the various analyses described in Table II.

Table III summarizes the means and standard deviations for each class section on each of the criterion variables and the two covariates. Adjusted means for the various covariate analyses are not given since there were a number of such analyses, each resulting in a different adjusted mean. There is no simple way to present all of these data and, in the interest of brevity, all have been omitted. Table III also shows the number of individuals in each class section. Numbers in parentheses represent the number of individuals for which complete data were available.

With over 24 different analyses of variance performed as a part of this study, it would consume a considerable amount of space to present all of the ANOVA tables in this paper. However, representative tables are presented. Tables IV-VII show the ANOVA tables for analyses 1(type Ia), 10(type IIb), 14(type IIIb), and 23(type IVa) respectively.

Table VIII summarizes the various analyses that were performed and the results that were obtained. By comparing the results of various analyses, some light is shed on the question of the importance of the correct choice of experimental unit in these analyses. During the discussion of this paper, attention will be focused on the following comparisons:

1 vs 2
3 vs 4

These analyses are of Type I (see Table II). The first number of each pair treats the individual as the experimental unit while the second member of the pair treats the class as the experimental unit. In the analyses in which the number of quizzes with scores of 8 or over (Q08) is used as the criterion, it is seen that there is no difference in the interpretation that would be made, regardless of the experimental unit. However, in analyses 3 and 4 where total points (TP) accumulated in the course is used as the criterion, it appears that use of the individual as the experimental unit may result in a conclusion that there is a significant

interaction between treatments R and O, whereas no significant differences are found when the division is used as the experimental unit.

These analyses differ from the previous ones in that a covariate has been added (Type II of Table II). In three of the four pairs, the conclusions that would be drawn when the individual is used as the experimental unit are identical to conclusions that would be drawn when the class is treated as the experimental unit. The exception is in analysis 6 where a significant main effect due to treatment R is indicated. It should be noted that the discrepancy between analyses 5 and 6 favors a significant difference in the case where the class is the experimental unit; the discrepancy between analyses 3 and 4 favors a significant difference in the case where the individual is the experimental unit.

5 vs 6

7 vs 8

9 vs 10

11 vs 12

These analyses have a third factor in the design but no covariate, (Type III in Table II). Once more, in three of the four comparisons, the results of the analyses are essentially the same. However, in analysis 13, a significant interaction is detected which does not show in analysis 14. The analysis which appears to be more powerful is the one which utilizes the individual as the experimental unit. Some readers may wish to question whether these comparisons are really parallel since one member of each pair involves stratification on the SAT score while the other member of the pair is a rather strange repeated measures analysis.

13 vs 14

15 vs 16

17 vs 18

19 vs 20

These analyses have both a third factor and a covariate in the design, (Type IV in Table II). The discrepancies in these analyses are the most distressing. In analysis 22, which uses the class as the experimental unit, a main effect which is significant at the .0003 level appears but there is no comparable effect seen in the analysis which utilizes the individual as the experimental unit. In spite of the very large F, the authors are inclined to believe that this is a spurious result. Although not likely due to chance, the difference is not likely due to treatment either. The explanation of the anomaly will be discussed.

21 vs 22

23 vs 24

Table III

Means and S.D. for each Section and Treatment Group

Section	N	Treatment	Total Points Mean	Total Points S.D.	SAT - V Mean	SAT - V S.D.	SAT - M Mean	SAT - M S.D.	# Quizzes rep. Mean	# Quizzes rep. S.D.	# Quizzes ≥8 Mean	# Quizzes ≥8 S.D.
1A	20(16)	rO	473.2	70.6	49.8	11.0	58.6	10.7	3.6	3.7	7.6	2.8
1B	20(18)	Ro	440.4	73.7	47.4	7.5	55.0	5.8	6.5	3.4	10.2	3.2
1C	18(14)	RO	466.4	98.3	45.6	7.9	56.7	7.9	5.8	3.0	11.0	3.8
2A	20(20)	rO	435.0	90.9	47.1	9.8	54.7	6.0	3.4	3.1	8.2	3.4
2B	17(13)	ro	482.5	96.7	48.8	9.2	57.1	5.6	4.3	3.3	9.7	3.8
3A	17(13)	ro	505.5	59.3	50.3	7.3	59.0	7.8	2.6	3.0	8.9	2.7
3B	16(10)	RO	444.6	50.1	45.5	7.3	53.6	8.9	6.8	3.4	11.4	2.7
4A	14(7)	Ro	507.1	66.4	50.1	7.1	56.7	8.5	7.0	4.0	10.0	3.6
4B	19(12)	rO	503.3	91.7	54.1	9.6	59.1	9.3	4.9	3.0	8.7	3.0
6A	18(14)	Ro	450.9	81.7	46.3	8.8	52.0	6.5	8.1	2.9	9.7	3.3
6B	22(18)	RO	490.2	72.3	50.9	9.6	59.3	7.1	6.1	3.4	10.9	3.2
6C	19(15)	ro	505.3	66.7	47.5	7.9	56.7	8.5	4.9	4.0	8.8	3.4
Total	220											
$\bar{X}$	18(14)											

TABLE IV

ANOVA Table for Analysis 1

Source of Variance	D.F.	S.S.	M.S.	F
Treatment R	1	147.221	147.221	14.376**
Treatment O	1	.147	.147	.014
R x O	1	43.465	43.465	4.244*
Residual	166	1699.936	10.241	
Total	169	1890.768		

** $p \leq .01$
* $p \leq .05$

TABLE V

ANOCVA Table for Analysis 10

Source of Variance	D.F.	S.S.	M.S.	F.
Treatment R	1	9.847	9.847	.033
Treatment O	1	1017.289	1017.289	3.441
R x O	1	409.339	409.339	1.385
Residual	6	1773.640	295.607	
Total	9	3210.115		

TABLE VI

ANOVA (REPEATED MEASURES)

TABLE for Analysis 14

Source of Variance	D.F.	S.S.	M.S.	F
Mean	1	8227278.7445	8227278.7445	10043.4870
Treatment R	1	1187.0322	1187.0322	.619
Treatment O	1	2445.9619	2445.9619	1.275
SAT-Math	2	28076.0950	14038.0475	17.1370**
R x O	1	502.6564	502.6564	.262
R x M	2	1768.5114	884.2557	1.0795
O x M	2	3891.2655	1945.6328	2.3751
U(RO)*	8	15350.2003	1918.7750	2.3424
ROM	2	1387.5300	693.7650	.8469
MU(RO)*	16	13106.6491	819.1656	

*Note that in the repeated measures with a mixed model, the estimate of the error variance is U(RO) for variables R and O but the best estimate is MU(RO) for variable M.

**$p \leq .01$

TABLE VII

ANOCOVA Table for Analysis 23

Source of Variance	D.F.	S.S.	M.S.	F
Treatment R	1	4.064	4.064	.547
Treatment O	1	4.410	4.410	.593
SAT-Math	2	19.480	9.740	1.310
R x O	1	66.760	66.760	8.979**
R x M	2	10.561	5.280	.710
O x M	2	13.161	6.580	.885
R x O x M	2	16.512	8.256	1.110
Residual	157	1167.361	7.435	
Total	168	1302.308		

**$p \leq .01$

TABLE VIII

Summary of Analyses and Results

Analysis Number	CODE: Criterion	CODE: Experimental Unit	CODE: 3rd Factor	CODE: Co-Variates	RESULTS
1.	QO8	SUB	---	---	R: F = 14.4; p = .0002 R x O: F = 4.2; p = .04
2.	QO8	DIV	---	---	R: F = 61.2; p = .0001 R x O: F = 18.9; p = .002
3.	TP	SUB	---	---	R x O: F = 3.9; p = .05
4.	TP	DIV	---	---	No Significant Differences
5.	QO8	SUB	---	Qrep	R x O: F = 8.6; p = .004
6.	QO8	DIV	---	Qrep	R: F = 7.8; p = .03 R x O: F 17.8; p = .004
7.	TP	SUB	---	V	R x O: F = 4.4; p = .04
8.	TP	DIV	---	V	R x O: F = 5.6; p = .05
9.	TP	SUB	---	VM	No Significant Differences
10.	TP	DIV	---	VM	No Significant Differences
11.	TP	SUB	---	M	No Significant Differences
12.	TP	DIV	---	M	No Significant Differences
13.	TP	SUB	M	---	M: F = 7.2; p = .007 R x O: F = 3.7; p = .05
14.	TP	DIV	M(rm)	---	M: F = 17.1; p = .0001
15.	TP	SUB	V	---	V: F = 13.4; p = .0008
16.	TP	DIV	V(rm)	---	V: F = 4.7; p = .02

TABLE VIII (continued)

17.	Q08	SUB	M	---	R: F = 12.7; p = .001 R x 0: F = 4.3; p = .04
18.	Q08	DIV	M(rm)	---	R: F = 31.1; p = .0005 R x 0: F = 12.3; p = .008
19.	Q08	SUB	V	---	R: F = 16.6; p = .001 R x 0: F = 4.0; p = .05
20.	Q08	DIV	V(rm)	---	R: F = 78.2; p = .0000? R x 0: F = 10.9; p = .01
21.	Q08	SUB	V	Qrep	R x 0: F = 8.5; p = .005
22.	Q08	DIV	V(rm)	Qrep	R: F = 35.3; p = .0003 R x 0: F = 12.9; p = .007
23.	Q08	SUB	M	Qrep	R x 0: F = 9.0; p = .005
24.	Q08	DIV	M(rm)	Qrep	R x 0: F = 10.3; p = .02

Legend

Q08 indicates that the criterion measure for this analysis was the number of quiz scores equal to or greater than eight.

TP indicates that the criterion measure for this analysis was the total number of points earned in the course.

SUB indicates that the individual subject was treated as the experimental unit in this analysis.

DIV indicates that the mean of the scores for individuals within a division (class section) was treated as the experimental unit in this analysis.

V represents the verbal score on the Scholastic Aptitude Test. When V is shown in the column headed "3rd Factor", it indicates that the sample was stratified into high, average, and low thirds on the basis of SAT-verbal score. When V appears in the column headed "Co-variates," it indicates that SAT-verbal scores were used as a covariate in the analysis.

M (rm) represents the mathematics score on the Scholastic Aptitude Test. Where this symbol follows V or M, it indicates that the analysis involved a repeated measures analysis rather than stratification on the variable.

Legend (continued)

Qrep refers to the number of quizzes which were repeated. (Students were allowed to repeat a quiz as many times as desired but the score recorded was the last score obtained.)

R represents the main effect of treatment R. (See page 4 for a description of the treatment.)

R x O represents an interaction between treatment R and treatment O. (See page 4 for a description of the treatments.)

Under the RESULTS column, each row represents a result which was statistically significant at the 0.05 level or beyond. In each row, the first letter represents the factor in the analysis which produced the significant F. This is followed by the value of F and the probability that an F of that value would occur by chance alone.

Sample Interpretation:

Refer to the row representing analysis number 24. In this analysis the number of quizzes with scores of eight or more was the criterion measure. The division (class section) was treated as the experimental unit, i.e. the "scores" treated in the statistical analysis were division means rather than individual scores. The analysis of variance utilized a 2 x 2 x 3 factorial design with treatments R and O as the first two factors. The "third factor" was a repeated measures using mean SAT-math scores for the high third, middle third, and low third of a division as the repeated measure. The number of quizzes repeated by the students were used as a covariate. This analysis produced no significant main effects. There was a significant R x O interaction which produced an F of 10.3. This value of F would occur by chance about 2 times in 100.